The feasibility of using low density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa

Aliloo, H; Mrode, R; Okeyo, AM; Ni, G; Goddard, ME; Gibson, JP

Published in:
Journal of Dairy Science

DOI:
10.3168/jds.2018-14621

Print publication: 01/10/2018

Document Version
Publisher's PDF, also known as Version of record

J. Dairy Sci. 101:9108–9127
https://doi.org/10.3168/jds.2018-14621

The feasibility of using low-density marker panels for genotype imputation and genomic prediction of crossbred dairy cattle of East Africa

H. Aliloo,*[1] R. Mrode,†‡ A. M. Okeyo,† G. Ni,* M. E. Goddard,§# and J. P. Gibson*
*School of Environmental and Rural Science, University of New England, Armidale, NSW 2350, Australia
†International Livestock Research Institute (ILRI), PO Box 30709, Nairobi, Kenya
‡Scotland's Rural College, Easter Bush, Midlothian EH25 9RG, Scotland, United Kingdom
§Agriculture Victoria, AgriBio, Centre for AgriBioscience, 5 Ring Road, Bundoora, VIC 3083, Australia
#Faculty of Veterinary and Agricultural Sciences, Department of Agriculture and Food Systems, The University of Melbourne, Parkville, VIC 3010, Australia

ABSTRACT

Cost-effective high-density (HD) genotypes of livestock species can be obtained by genotyping a proportion of the population using a HD panel and the remainder using a cheaper low-density panel, and then imputing the missing genotypes that are not directly assayed in the low-density panel. The efficacy of genotype imputation can largely be affected by the structure and history of the specific target population and it should be checked before incorporating imputation in routine genotyping practices. Here, we investigated the efficacy of imputation in crossbred dairy cattle populations of East Africa using 4 different commercial single nucleotide polymorphisms (SNP) panels, 3 reference populations, and 3 imputation algorithms. We found that Minimac and a reference population, which included a mixture of crossbred and ancestral purebred animals, provided the highest imputation accuracy compared with other scenarios of imputation. The accuracies of imputation, measured as the correlation between real and imputed genotypes averaged across SNP, were around 0.76 and 0.94 for 7K and 40K SNP, respectively, when imputed up to a 770K panel. We also presented a method to maximize the imputation accuracy of low-density panels, which relies on the pairwise (co)variances between SNP and the minor allele frequency of SNP. The performance of the developed method was tested in a 5-fold cross-validation process where various densities of SNP were selected using the (co)variance method and also by alternative SNP selection methods and then imputed up to the HD panel. The (co)variance method provided the highest imputation accuracies at almost all marker densities, with accuracies being up to 0.19 higher than the random selection of SNP. The accuracies of imputation from 7K and 40K panels selected using the (co)variance method were around 0.80 and 0.94, respectively. The presented method also achieved higher accuracy of genomic prediction at lower densities of selected SNP. The squared correlation between genomic breeding values estimated using imputed genotypes and those from the real 770K HD panel was 0.95 when the accuracy of imputation was 0.64. The presented method for SNP selection is straightforward in its application and can ensure high accuracies in genotype imputation of crossbred dairy populations in East Africa.

Key words: genotype imputation, genomic selection, low-density marker panel design, East African crossbred dairy cattle

INTRODUCTION

Selection of animals based on genomic estimated breeding values (**GEBV**); that is, genomic selection (**GS**), is now a standard practice for genetic improvement of many livestock species. Genomic selection exploits the linkage disequilibrium (**LD**) between known markers and unknown causal mutations in estimation of GEBV. Genome-wide SNP are usually used as genomic markers to estimate GEBV of selection candidates that have genotypes only, based on a prediction equation that is derived from a large reference population with both genotypes and phenotypes (Meuwissen et al., 2016).

Genomic selection is especially important in situations where traditional genetic evaluations based on pedigrees are not available because of the absence of pedigree information. Smallholder dairy farmers in East Africa rear crossbred cattle to combine the adaptation features of indigenous animals with the high milk yield potential of exotic dairy breeds. These farmers do not record pedigrees and there is no current genetic evaluation to aid them in making informed breeding decisions. Based on new phenotype recording programs, GS could help East African smallholder dairy farmers to establish an effective genetic improvement program.

Received February 22, 2018.
Accepted May 26, 2018.
[1]Corresponding author: haliloo@une.edu.au

The accuracy of GEBV can increase when a larger reference population and high-density (**HD**) SNP panels are used for estimation of marker effects in the reference population. Denser panels are more effective in capturing LD between markers and QTL, and although medium density 50K assays are dense enough for reaching a useful level of LD within a breed, HD panels are required for multi-breed applications (e.g., de Roos et al., 2008). This is especially important in situations where the size of reference population for a breed is small and genotypes from other larger breeds are incorporated in genomic prediction. Although the cost of genotyping has decreased dramatically since the technology emerged, HD SNP panels are still very costly for routine use in genetic improvement of livestock species, especially in smallholder systems. A cost-effective alternative is to genotype animals with cheaper low density panels and then to infer the missing genotypes that have not been directly assayed, based on information from a reference population genotyped by an HD panel; a method called genotype imputation.

The optimal number of SNP and an appropriate algorithm for selecting them to include in a low-density array that can later be used in imputation to HD genotypes is unknown. Habier et al. (2009) suggested to genotype selection candidates using a sparse panel of evenly spaced marker across the genome and then impute the missing genotypes using co-segregation information within families. Although their proposed method could work across different breeds and traits and is independent of the genetic architecture of trait of interest, it requires availability of pedigree and HD genotypes on both parents of selection candidates. Other attempts to design low-density SNP panels has been mostly based on the use of evenly spaced markers and maximization of minor allele frequency (**MAF**) with some enrichments at chromosomal ends (e.g., Boichard et al., 2012; Bolormaa et al., 2015). Corbin et al. (2014) showed that when the low-density panels are designed to optimize equidistant spacing of markers based on LD units and to increase MAF, they can provide higher imputation accuracy and lower variations in accuracy of individual SNP than equidistant selection of SNP on base pair positions. Wu et al. (2016) described a multiple objective optimization algorithm to select SNP for low-density panels that achieved substantially higher imputation accuracies than when selecting SNP solely based on uniform distribution of map position.

Knowledge on the level and extent of LD between genome-wide markers is important because it can help to determine the required number of SNP markers for fine mapping of quantitative trait loci, GS, and genotype imputation (e.g., Sargolzaei et al., 2008; Corbin et al., 2014; Mathew et al., 2018). The structure of LD is different in different populations. It is expected that in populations with smaller effective population size (N_e) and higher average LD between markers, such as commercial dairy cattle breeds, lower number of markers will suffice. It has also been suggested that HD panels with at least 300,000 SNP are required for multi-breed applications (de Roos et al., 2008).

Existing SNP assays have been mainly designed for use in pure breeds and methods of imputation have been tested mostly in purebred populations. East African crossbred dairy cattle populations are complex admixtures of dairy *Bos taurus* breeds and indigenous African breeds. Therefore, the objectives of this study were to assess the accuracy of genotype imputation and subsequent genomic prediction in crossbred dairy cattle populations of East Africa. We compared existing arrays and methods of imputation and various methods of selecting SNP for customized arrays, including a new method based on (co)variances between SNP that are weighted by their MAF.

MATERIALS AND METHODS

Data

Population. The crossbred dairy cattle in East Africa form an admixed population resulting from many generations of crossing of African indigenous cattle to several exotic dairy breeds, mainly from Friesian, Holstein, Ayrshire and related red breeds, and Jersey. These animals are kept by smallholder dairy farmers, typically in herds of size 1 to 5 cows, and produce almost all of the milk consumed in East Africa. The majority of East African crossbred dairy cattle are bred via natural mating, with a small proportion of matings by AI to imported and locally bred purebred dairy bulls. Very few animals have pedigree records and no genetic evaluation systems or systematic breeding programs are used to aid farmers. The Dairy Genetics East Africa (**DGEA**) project collected a wide range of smallholder cow performance, animal genotype, and household data in 4 east African countries, Kenya, Uganda, Ethiopia, and Tanzania, between 2010 and 2014 to determine the needs and provide feasible solutions for short- and long-term genetic improvement of smallholder crossbred dairy cattle populations.

The genetic diversity of the crossbred cattle in relation to the indigenous breeds of the region and global reference breeds was previously presented in principal component plots by Strucken et al. (2017). They showed that the East African indigenous breeds are ancient admixtures of *Bos indicus* and African *Bos taurus* cattle where the latter is a lineage that is genetically very distinct from European *Bos taurus*. The crossbred dairy

cattle were all clearly shown in the principal component plots as crosses between exotic dairy breeds and the local indigenous breeds of the country in which they were sampled, with proportion of exotic dairy content ranging from almost 0 to almost 100% (Strucken et al., 2017).

Genotypes. Genotype data were available on 3,513 animals (3,124 crossbreds and 389 indigenous breed animals) genotyped for 777,962 SNP markers using the Illumina BovineHD BeadChip (Illumina, San Diego, CA). Animals consisted of indigenous breeds from East Africa as well as crossbred cows (from Kenya, Uganda, Ethiopia, and Tanzania) and bulls (only from Kenya and Uganda) in those countries. Another data set containing 26 British Friesian and 519 Canadian Ayrshire cows genotyped on the same SNP panel was also added to the data to increase the size of purebred genotypes. Quality controls applied on the combined raw data were as follows: only SNP with GC score >0.6 and call rate >95% were kept; mitochondrial, unmapped, duplicate map position, and SNP located on sex chromosomes (X and Y) were removed. Further, SNP with a MAF less than 0.01 were excluded. Animals were also required to have genotypes for more than 90% of SNP. These controls resulted to 691,230 SNP over 29 *Bos taurus* autosomes based on UMD 3.1 genome assembly (Zimin et al., 2009). To increase the size of data further and to have more purebred animals that could be used as the reference population for genotype imputation, 197 animals representing Holstein, Jersey, Guernsey, Nelore, and N'Dama breeds genotyped by the bovine HapMap consortium (http://bovinegenome.org) were added to the data. The publicly available HapMap genotypes were post quality control. Only those SNP in common between the HapMap and DGEA genotypes (after quality control) were included. The 5 dairy breeds included in the data set represented the main dairy breeds reported to have been used for crossbreeding in the region. Finally, there were 4,207 animals whose SNP marker genotypes were coded as 0, 1, and 2, respectively, for AA, AB, and BB allele combinations. Table 1 contains the details on the number of animals in different breeds and the sources of data for this study.

After quality controls, 0.58% of genotypes were sporadically missing. To have a complete data set for all animals at all loci, these sporadically missing genotypes of individuals were imputed using FImpute V 2.2 (Sargolzaei et al., 2014). To test the accuracy of this imputation, we randomly masked genotypes for 5% of the known genotypes and then all the missing genotypes (~5.58%) were predicted together by FImpute.

Table 1. Number of genotyped animals for different breeds and breed groups and the sources they were obtained from

Breed[1]	Size	Source[2]	Breed group	Breed group size
Ethiopian Begait Barka	30	DGEA	African Zebu	285
Ethiopian Boran	28	DGEA	African Zebu	
Ethiopian Central Highland	28	DGEA	African Zebu	
Ethiopian Danakil Harar	30	DGEA	African Zebu	
Ethiopian Fogera	28	DGEA	African Zebu	
Kenyan Boran	28	DGEA	African Zebu	
Kenyan Zebu	58	DGEA	African Zebu	
Tanzanian Boran	20	DGEA	African Zebu	
Tanzanian Iringa Red	13	DGEA	African Zebu	
Tanzanian Singida White	22	DGEA	African Zebu	
Ugandan Ankole	43	DGEA	Sanga Zebu	60
Ugandan Nganda	17	DGEA	Sanga Zebu	
Ethiopian crossbred	545	DGEA	African Crossbred	3,083
Kenyan crossbred bull	97	DGEA	African Crossbred	
Kenyan crossbred	1,378	DGEA	African Crossbred	
Tanzanian crossbred	462	DGEA	African Crossbred	
Ugandan crossbred bull	46	DGEA	African Crossbred	
Ugandan crossbred	555	DGEA	African Crossbred	
Kenyan Sahiwal	38	DGEA	*Bos indicus*	73
Nelore	35	HapMap	*Bos indicus*	
N'Dama	24	HapMap	African *taurine*	24
Guernsey	21	HapMap	*Bos taurus*	682
Holstein	71	HapMap	*Bos taurus*	
Jersey	46	HapMap	*Bos taurus*	
British Friesian	25	SRUC	*Bos taurus*	
Ayrshire	519	CDN	*Bos taurus*	

[1]All animals were females except 2 crossbred male populations from Kenya and Uganda.

[2]DGEA = Dairy Genetics East Africa; HapMap = the Bovine HapMap consortium; SRUC = Scottish Rural University College; and CDN = Canadian Dairy Network.

The correlation between imputed genotypes and the masked genotypes (5%) were higher than 99%, indicating a very high accuracy of imputation.

To explore the genetic diversity in different breeds under investigation, LD was computed for phased genotypes of animals as the squared correlation coefficient of haplotypes of syntenic loci that were up to 1 Mbp apart:

$$r^2 = \frac{\left(p_{AB} - p_A p_B\right)^2}{p_A p_a p_B p_b},$$

where A and a, and B and b, are alleles of A and B SNP, respectively; p_A, p_a, p_B, and p_b are the corresponding allele frequencies; and p_{AB} is the frequency of the AB haplotype.

Phenotypes. Test-day milk yield records (**TDMY**) were available from the first 3 parities of 1,034 smallholder crossbred cows aged between 4 and 8 yr in Kenya. A fixed regression test-day model was used fitting contemporary group effects of random management group-year-season and fixed parity. Fixed lactation curves of animals were modeled by Legendre polynomials of order 4 within days in milk interacting with dairy group. Animals were assigned into 5 dairy groups based on their percentage dairy breed ancestry estimated from an admixture analysis (Ojango et al., 2014). The model also included a random additive animal effect, $N\left(\mathbf{0}, \mathbf{G}\sigma_a^2\right)$, where **G** is the additive relationship matrix (described later), plus a permanent environmental effect, $N\left(\mathbf{0}, \mathbf{I}\sigma_{pe}^2\right)$, with **I** being the identity matrix. Milk yield deviations (**MYD**) were calculated by correcting TDMY for fixed effects, management group-year-season effects, and permanent environmental effects estimated from the fixed regression test-day model. For each animal a single MYD was calculated as the average of all MYD of the animal. Different animals had different number of TDMY and to account for the effect of the number of TDMY on the accuracy of the calculated MYD, a weight was assigned to each MYD based on the inverse of the standard error of each MYD. The standard error of each MYD was calculated as the standard deviation of MYD divided by the square root of n, with n being the number of TDMY used for calculation of MYD.

SNP Selection

To design lower density SNP panels that can be efficiently used for genotype imputation to higher densities or used directly in genetic evaluation of genotyped animals, a method of selecting SNP based on the pairwise SNP (co)variance was developed. Consider n SNP from which we want to select k SNP such that the selected k SNP together explain a higher proportion of the variance of the n SNP than any other set of k SNP. To start the SNP selection process, SNP genotypes are scaled so that the mean and variance at each SNP are 0 and 1, respectively:

$$x_{k_{adj}} = \frac{\left(x_k - \overline{x_k}\right)}{\sigma\left(x_k\right)},$$

where x_k is genotype at kth SNP and $\overline{x_k}$ and $\sigma\left(x_k\right)$ are the average and standard deviation of kth SNP genotype, respectively.

Then the covariances between all pairs of scaled SNP genotypes are calculated and stored in a matrix (**V**), which is an $n \times n$ (co)variance matrix, and V_{ij} is the covariance between SNP i and SNP j. The diagonal elements of matrix **V** are variances of SNP, and initially all are equal to 1. The sum of the diagonal elements or the trace of **V** is the total variance of n SNP, which is equal to the total number of SNP in the beginning.

The developed method for SNP selection (COV) is a sequential process where in each round:

(1) For a given SNP the strength of its correlation with all other SNP is calculated. This is summed up across all pairs for each SNP and is weighted by the MAF of the given SNP:

$$i = 1,\ldots,n;\ j = 1,\ldots,n \text{ and } i \neq j,$$

$$E_{ij} = V_{ii} - \frac{V_{ij}^2}{V_{jj}}, \text{and}$$

$$D_j = \sum_{i=1}^{n} E_{ij}, \text{and}$$

$$D_{j_{adj}} = D_j \times \left[1 - \left(w \times \text{MAF}_j\right)\right],$$

where E_{ij} is the unexplained variance (UNV) for SNP i after accounting for SNP j; D_j is the sum of UNV across all SNP for SNP j, and w is the weight on the MAF, which can be between 0 and 1. If $w = 0$, MAF is ignored in selection of SNP, whereas with $w = 1$ the same weight is put on both UNV and MAF. We used a weighing factor $w = 1$ in the current study.

The SNP with the lowest adjusted D_{adj}, say SNP k, is selected because it has highest average

covariance with all the other SNP, so it explains more variance than any other SNP and it is also highly informative because of being highly polymorphic.

(2) The pairwise (co)variances between the remaining SNP are corrected by removing the amount of (co)variance explained by covariance of each SNP with the selected SNP, k:

$$i = 1,\ldots,n-1 \text{ and } j = 1,\ldots,n-1,$$

$$V_{ij_{adj}} = V_{ij} - \frac{V_{ik} \times V_{jk}}{V_{kk}},$$

where $V_{ij_{adj}}$ is the (co)variance between SNP i and SNP j corrected for the selected SNP k. Then, the adjusted (co)variance matrix $\mathbf{V}_{adj}$ has dimensions of $(n-1) \times (n-1)$. The SNP in perfect LD with the selected SNP will have the same D value, and therefore they are removed at this stage because 100% of the information they contributed is already explained by the selected SNP.

(3) At this stage, it is determined whether the selected SNP have explained enough variance and the SNP selection process should be stopped or more SNP are required. The proportion of variance explained by the selected SNP is calculated as

$$\omega^2_{exp_t} = \frac{\left(\sigma^2_0 - \sigma^2_t\right)}{\sigma^2_0},$$

where ω^2_{expt} is the proportion of total variance explained by selected SNP after selecting t SNP, σ^2_0 is the total variance with no SNP selected and σ^2_t is the remaining variance after t SNP selected and is calculated as the trace of $\mathbf{V}_{adj}$.

We used a sliding window approach in which SNP were selected within overlapping intervals of 1 Mbp. The interval moved forward by 500 kbp until it reached the end of the chromosome. The number of SNP selected from each window was determined based on the proportion of variance that was required to be explained by the selected SNP. Different thresholds for the proportion of explained variance $\left(\omega^2_{exp}\right)$ were set to achieve different densities of SNP panels. To account for the edge effect, twice the number of SNP required for explaining variance were selected from the first and last 1 Mbp interval in each chromosome. We also selected equal number of SNP to that selected by the (co)variance method (**COV**) within each interval either based on highest MAF (**MAFI**) or randomly (**RANI**). The SNP were also selected randomly (**RANC**) or based on highest MAF (**MAFC**) across the whole chromosome without accounting for their map position on the chromosome, matching the number of SNP selected by the (co)variance method at each density.

The SNP selection was carried out using only crossbred animals. To implement the SNP selection and validation in independent populations, a cross-validation approach was implemented for the COV, MAFI, and MAFC methods. Animals were randomly divided into 5 groups such that the number of animals in each group was as similar as possible and animals from all countries are presented in each group (~617 animals in each fold). Then at each rotation, 4 folds were used to select SNP and 1 fold was excluded from SNP selection and was only used in the validation processes (genomic prediction and imputation). The random selections of SNP within interval (RANI) or across chromosome (RANC) were also repeated 5 times to minimize the sampling error.

To assess the efficiency of the developed method for selecting SNP, the selected SNP by the 5 different methods were used in turn for genome-enabled best linear unbiased prediction (**GBLUP**) of breeding values of animals with both genotypes and phenotypes (n = 1,034) and for genotype imputation of crossbred animals to the HD panel (details below). The imputation and GBLUP accuracies obtained from SNP selected by the 5 selection methods were averaged across the 5 folds.

Genotype Imputation

Four different commercially available SNP chips, Illumina BovineLD v2 (m = 7,931), BovineSNP50 v3 BeadChip (m = 53,218; Illumina), GeneSeek-GenomicProfiler (**GGP**) Bovine 50K (m = 47,843), and Indicus 35K v1.03 (m = 34,000; Neogen Corporation, Lincoln, NE), with m being the number of SNP in the original panel, were used to find the optimal strategy for genotype imputation. Different scenarios of imputation in which different groups of animals were included in the reference population to predict the genotypes of crossbred animals were tested. The reference population consisted of only crossbred (scenario 1), only purebred (scenario 2), or both crossbred and purebred animals (scenario 3). At each rotation of cross-validation in scenario 1, the reference and validation sets included around 2,466 and 617 animals, respectively. Scenario

2 used 1,124 reference animals to impute 3,083 crossbreds. The reference set in scenario 3 was around 3,590 at each rotation of cross-validation to impute around 617 animals. Further, we also investigated whether the choice of imputation algorithm can affect the imputation accuracy. Three different programs, FImpute v2.2 (Sargolzaei et al., 2014), Beagle v4.1 (Browning and Browning, 2016), and Minimac v3 (Das et al., 2016), were used as the choice for imputation software. For FImpute and Beagle, genotype phasing was done using their embedded algorithms during imputation, whereas for Minimac, Eagle v2.3.5 (Loh et al., 2016) was used for pre-phasing genotypes before imputation. The SNP in common between each of the 4 commercial panels and the HD panel were extracted and then used in imputation to the HD panel. The best imputation strategy was defined as the scenario that resulted in the highest imputation accuracy. The accuracy of imputation was measured as the proportion of correctly imputed genotypes (i.e., concordance), as well as the correlation between real and imputed genotypes averaged across SNP. The optimal imputation strategy was then used for imputing up the SNP selected by the different SNP selection methods to the HD panel.

Relationship Between Animals in Target and Reference Set

To investigate the effect of connectedness between training and target populations on imputation accuracy, various measures of genomic relationship between animals in reference and validation sets were calculated for different scenarios of imputation. For each animal in the target set, we calculated the maximum, average of top 5 and top 10, as well as all genomic relationship coefficients between that animal and animals from the reference set that were used in the imputation of the given animal. Further, we also calculated an average value across all validation animals for all replications of cross-validation to compare the imputation scenarios in terms of relationships between reference and target sets. For these comparisons, the allelic frequency was set to 0.5 for all SNP genotypes so that the difference in allele frequencies between breeds did not affect the relationships and the genomic relationships were only derived by the differences in actual genotypes.

Genomic Prediction

The GEBV of animals with both genotypes and phenotypes (n = 1034) were estimated using a linear mixed model in GBLUP context:

$$\mathbf{y} = \mathbf{1_n}\mu + \mathbf{Wu} + \mathbf{e},$$

where $\mathbf{y}$ is the vector of MYD; $\mathbf{1_n}$ is a vector of ones; μ is the population mean term; $\mathbf{u}$ is a vector that contains genomic breeding values of animals and is assumed to be distributed as $\mathbf{u} \sim N\left(\mathbf{0}, \mathbf{G}\sigma_a^2\right)$, where $\mathbf{G}$ is the additive relationship matrix based on SNP genotype; $\mathbf{e}$ is the vector of random residual term distributed as $\mathbf{e} \sim N\left(\mathbf{0}, \mathbf{S}\sigma_e^2\right)$, $\mathbf{S}$ being a diagonal matrix with weight values for each MYD; and $\mathbf{W}$ is the incidence matrix relating phenotypes to animals. $\mathbf{G}$ was constructed according to VanRaden (2008):

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{\sum_{k=1}^{m} 2p_k q_k},$$

where $\mathbf{Z}$ is the matrix for additive marker covariates and contains $0 - 2p_k$, $1 - 2p_k$, and $2 - 2p_k$ for AA, AB, and BB genotypes, respectively; p_k is the frequency of allele B at marker k and $q_k = 1 - p_k$.

The GEBV were computed for animals using the HD panel ($\mathbf{GEBV_{HD}}$) in a 5-fold cross-validation process treating 1 fold as the target set and the rest as reference at each rotation. The GEBV were also estimated using different reduced subsets of selected SNP attained by various methods ($\mathbf{GEBV_{SEL}}$) as well as using imputed genotypes to the HD panel ($\mathbf{GEBV_{IMP}}$) in the same cross-validation setting. The correlations between $GEBV_{HD}$ and $GEBV_{SEL}$ and those between $GEBV_{HD}$ and $GEBV_{IMP}$ were calculated and averaged across 5 folds to assess the performance of different SNP selection methods.

RESULTS

Genetic Diversity

Figure 1 illustrates the decay in pairwise LD between SNP located at varying distances on genome for crossbreds compared with purebred populations. As expected, LD is higher between SNP in close proximity and it decreases as the distance between SNP increases. The average LD was found to be lower in crossbreds compared with purebred animals at all distances. In addition, LD in crossbreds shows a more rapid decline over distance than those in purebreds (Figure 1). Among the exotic purebred populations, Guernsey and Jersey had stronger LD between pairwise markers at all distances and British Friesians showed lowest levels of LD. Iringa Red and Kenyan Boran showed substantially higher LD

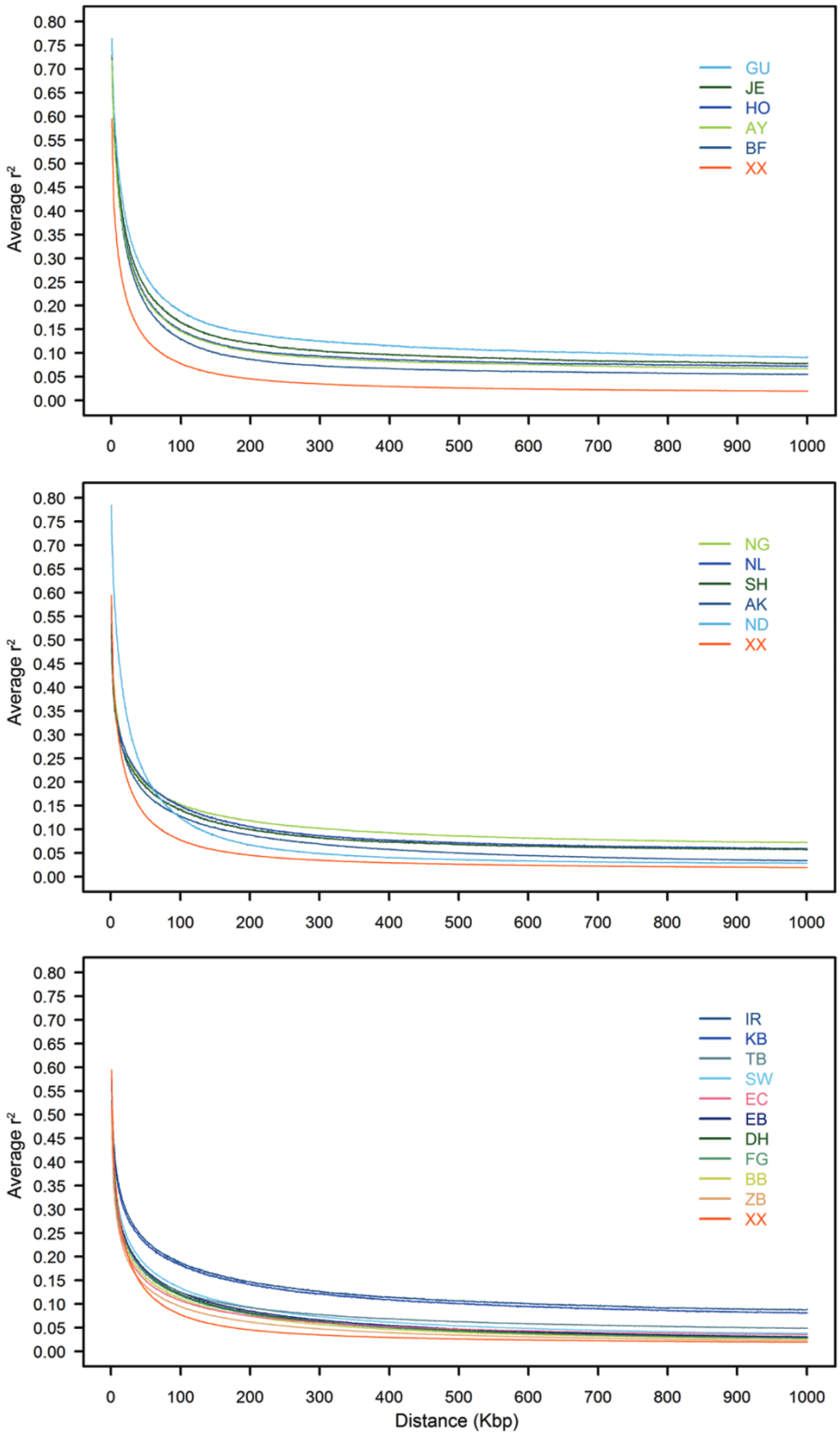

Figure 1. Average linkage disequilibrium (r^2) of pairwise SNP over varying genomic distances in crossbred (XX) versus (top) 5 *Bos taurus* (GU = Guernsey; JE = Jersey; HO = Holstein; AY = Ayrshire; and BF = British Friesian) breeds; (middle) 2 *Bos indicus* (NL = Nelore and SH = Sahiwal), 2 Sanga (NG = Nganda and AK = Ankole), and 1 African *taurine* (ND = N'Dama) breeds; and (bottom) 10 African Zebu (IR = Iringa Red; KB = Kenyan Boran; TB = Tanzanian Boran; SW = Singida White; EC = Central Highland; EB = Ethiopian Boran; DH = Danakil Harar; FG = Fogera; BB = Begait Barka; and ZB = Zebu) breeds.

than other indigenous Zebu breeds, which were similar to the levels of LD that were observed in Nganda.

Optimal Imputation Strategy

The total number of SNP in the 4 commercially available arrays retained from the HD panel after quality control as well as the number of common SNP between the different panels are included in Table 2. Illumina Bovine50 had the highest number of SNP in common with the HD panel, followed by the GGP Bovine 50K. The GGP Indicus 35K had the lowest number of SNP in common with all the other panels, reflecting the selection of SNP with high MAF in *Bos indicus* on the GGP Indicus 35K. The accuracy of imputation of crossbred genotypes obtained from the 3 different imputation algorithms are shown in Tables 3 and 4. As expected, marker panels with higher number of SNP generally achieved higher imputation accuracies in all scenarios of imputation. However, the GGP Bovine 50K always achieved higher imputation accuracies than the Illumina Bovine50 though it contained around 2.6K fewer SNP. The difference between imputation accuracy of the GGP Bovine 50K and the Illumina Bovine50 was higher when Beagle was used as the imputation software. For FImpute and Minimac, most of the accuracy in imputation of crossbred genotypes came from including crossbred animals in the reference data (scenario 1) and putting purebreds in the reference set (scenario 3) added little accuracy. This was not the case in imputations carried out by Beagle where scenario 3 achieved up to 0.11 higher accuracy than scenario 1 (Table 4). Beagle achieved lowest imputation accuracy in scenario 1 compared with other imputation algorithms with higher differences for low-density panels. Scenario 2 resulted in the lowest imputation accuracies for all commercial panels and imputation programs whereas scenario 3 always achieved the highest imputation accuracies. Among the 3 imputation software programs used in this study, Minimac outperformed FImpute and they both performed better than Beagle. The difference between imputation software programs was especially higher in imputations of lower densities such that in scenario 3 Minimac achieved up to 0.33 and 0.15 higher correlations in imputation of Illumina BovineLD than those from Beagle and FImpute, respectively (Tables 4).

The combination of Minimac and scenario 3 was chosen as the optimal imputation strategy because it provided the highest imputation accuracy for all SNP panels compared with other imputation algorithms and scenarios. Figure 2 shows the concordance and correlation of imputations for different chromosomes obtained from the optimal imputation strategy. Accuracy of individual chromosomes was similar to the overall imputation accuracies reported for different panels (Tables 3 and 4). However, chromosomal accuracies had fewer fluctuations when the overall imputation accuracy was higher. Different chromosomes achieved the highest and lowest imputation accuracies in different panels. For example, chromosomes 5, 6, 2, and 8 showed the highest concordance in imputation of GGP Bovine 50K, Illumina Bovine50, GGP Indicus 35K, and Illumina BovineLD, respectively. While correlation and concordance showed very similar trends across different chromosomes, correlations were lower than concordances especially for Illumina BovineLD (Figure 2).

The values of concordance and correlations of imputed SNP against their MAF are illustrated in Figure 3. For all panels, SNP with lowest MAF showed higher concordance values and concordance decreased as MAF increased. The rate of decline in concordance values was highest for Illumina BovineLD compared with other panels, and the panel with highest average concordance across all SNP (i.e., GGP Bovine 50K) showed smaller reduction in SNP concordance values. In contrast, correlations showed a moderate ascending trend from the low to high MAF and also less variations across MAF for all panels.

Reference and Target Population Connectedness

Table 5 contains various measures of genomic relationships used to evaluate the connectedness between reference and target animals in different scenarios of imputation. The values of connectedness across the

Table 2. Total number of SNP (diagonals) and number of common SNP between different commercial panels (lower triangle) retained from the quality controlled high-density genotypes (with proportion of missing SNP to be imputed in parentheses)

Panel[1]	Illumina BovineLD	GGP Indicus 35K	GGP Bovine 50K	Illumina Bovine50
Illumina BovineLD	7,154 (0.99)			
GGP Indicus 35K	1,252	30,586 (0.96)		
GGP Bovine 50K	6,998	3,532	39,480 (0.94)	
Illumina Bovine50	7,063	2,345	13,873	42,147 (0.94)

[1]GeneSeek-Genomic-Profiler (GGP) Bovine 50K and GGP Indicus 35K (Neogen Corporation, Lincoln, NE); Illumina Bovine50 and Illumina BovineLD (Illumina, San Diego, CA).

Table 3. Concordance[1] values for imputation of commercial panels in different scenarios of imputation from Beagle, FImpute, and Minimac (with SE in parentheses)

		Imputation scenario[4]		
Panel[2]	Software[3]	1	2	3
Illumina BovineLD	Beagle	0.6175 (0.0004)	0.6001	0.6222 (0.0001)
	FImpute	0.7766 (0.0008)	0.6437	0.7779 (0.0007)
	Minimac	0.8556 (0.0006)	0.7072	0.8564 (0.0006)
GGP Indicus 35K	Beagle	0.7753 (0.0001)	0.6946	0.8531 (0.0002)
	FImpute	0.9141 (0.0004)	0.8077	0.9175 (0.0004)
	Minimac	0.9483 (0.0003)	0.8644	0.9519 (0.0002)
GGP Bovine 50K	Beagle	0.8837 (0.0004)	0.7367	0.9456 (0.0003)
	FImpute	0.9366 (0.0004)	0.8519	0.9386 (0.0004)
	Minimac	0.9564 (0.0003)	0.8921	0.9597 (0.0003)
Illumina Bovine50	Beagle	0.8456 (0.0004)	0.7238	0.9151 (0.0003)
	FImpute	0.9274 (0.0004)	0.8387	0.9295 (0.0004)
	Minimac	0.9499 (0.0004)	0.8812	0.9534 (0.0003)

[1]Concordance was defined as the proportion of correctly imputed genotypes.

[2]GeneSeek-Genomic-Profiler (GGP) Bovine 50K and GGP Indicus 35K (Neogen Corporation, Lincoln, NE); Illumina Bovine50 and Illumina BovineLD (Illumina, San Diego, CA).

[3]FImpute v2.2 (Sargolzaei et al., 2014), Beagle v4.1 (Browning and Browning, 2016), and Minimac v3 (Das et al., 2016).

[4]Imputation scenarios differed based on the inclusion of animals in the reference population where in scenarios (1) only crossbred, (2) only purebred and (3) all purebred and crossbred were included in the reference set.

3 scenarios agreed with the imputation accuracies obtained from each scenario (Tables 3 and 4), where higher connectedness always led to higher imputation accuracy.

Figures 4 and 5 show the mean values of concordance and correlations, respectively, for target animals grouped according to their average genomic relationship with the reference set used in imputation of their genotypes. For all panels and in all scenarios of imputation, concordance and correlation increased as the relationship between target and reference sets. The accuracies of the imputation from Illumina BovineLD to HD genotypes benefitted the most from the increase in genomic relationships between target and reference animals. For example in scenario 3, the difference between concordance values of animals with the lowest and highest genomic relationship with reference set was around 0.26 for imputation from Illumina BovineLD

Table 4. Correlations[1] for imputation of commercial panels in different scenarios of imputation from Beagle, FImpute, and Minimac (with SE in parentheses)

		Imputation scenario[4]		
Panel[2]	Software[3]	1	2	3
Illumina BovineLD	Beagle	0.4501 (0.0003)	0.3709	0.4344 (0.0002)
	FImpute	0.6136 (0.0012)	0.3657	0.6177 (0.0012)
	Minimac	0.7637 (0.0007)	0.4734	0.7638 (0.0007)
GGP Indicus 35K	Beagle	0.6384 (0.0003)	0.5692	0.7569 (0.0003)
	FImpute	0.8630 (0.0006)	0.6825	0.8688 (0.0006)
	Minimac	0.9217 (0.0004)	0.7847	0.9274 (0.0003)
GGP Bovine 50K	Beagle	0.7958 (0.0003)	0.6204	0.9078 (0.0004)
	FImpute	0.8944 (0.0004)	0.7500	0.8979 (0.0004)
	Minimac	0.9296 (0.0004)	0.8227	0.9350 (0.0003)
Illumina Bovine50	Beagle	0.7516 (0.0003)	0.6026	0.8587 (0.0003)
	FImpute	0.8822 (0.0004)	0.7356	0.8858 (0.0004)
	Minimac	0.9209 (0.0004)	0.8104	0.9265 (0.0004)

[1]Correlations between masked and imputed genotypes averaged across SNP.

[2]GeneSeek-Genomic-Profiler (GGP) Bovine 50K and GGP Indicus 35K (Neogen Corporation, Lincoln, NE); Illumina Bovine50 and Illumina BovineLD (Illumina, San Diego, CA).

[3]FImpute v2.2 (Sargolzaei et al., 2014), Beagle v4.1 (Browning and Browning, 2016), and Minimac v3 (Das et al., 2016).

[4]Imputation scenarios differed based on the inclusion of animals in the reference population where in scenarios (1) only crossbred, (2) only purebred, and (3) all purebred and crossbred were included in the reference set.

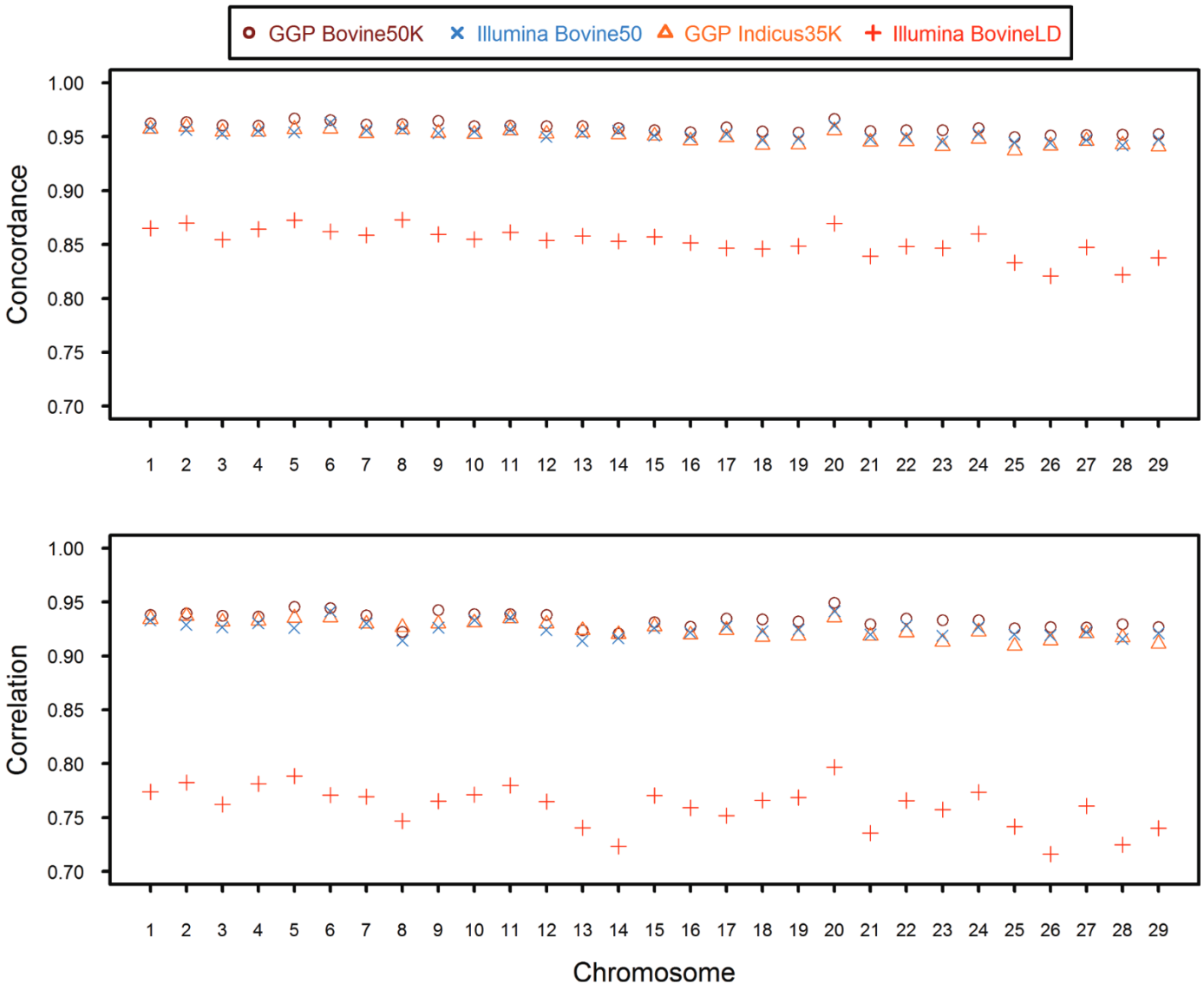

Figure 2. Concordance values (top) and correlations (bottom) of individual chromosomes obtained from the optimal imputation of different commercial arrays. GeneSeek-Genomic-Profiler (GGP) Bovine 50K and GGP Indicus 35K (Neogen Corporation, Lincoln, NE); Illumina Bovine50 and Illumina BovineLD (Illumina, San Diego, CA). Color version available online.

whereas other panels showed an average difference of 0.15.

Imputation Accuracies from Different SNP Selection Methods

Table 6 shows the number of selected SNP at different thresholds for the total variance of SNP explained, as well as the accuracies of imputation obtained from different SNP selection methods. Selection of SNP based on the (co)variance method achieved the highest imputation accuracy at almost all thresholds such that it provided 0.03, 0.19, 0.17, and 0.16 higher correlations compared with SNP selection based on MAFI, RANI, RANC, and MAFC, respectively, when around 4K SNP were selected (5% of total SNP variance explained). The values of concordance and correlations within a SNP selection method could differ to a large extent especially at lower densities of selected SNP, but they became closer as more SNP were selected at higher densities. The difference between the accuracy of imputation from the (co)variance method and those of other SNP selection methods was also highest at lower marker densities and there was little difference in accuracy of imputation between methods at high marker densities, where all accuracies were high. Selection of

Table 5. Connectedness between animals in target and reference sets measured by genomic relationships in different scenarios of imputation

Connectedness measure[1]	Scenario		
	1	2	3
Max	0.73	0.65	0.73
Top 5	0.67	0.65	0.68
Top 10	0.65	0.65	0.67
Mean	0.55	0.53	0.55

[1]Max = average of maximum relationships; top 5 = average of top 5 relationships; top 10 = average of top 10 relationships; and mean = average of all relationships between each individual in the target set and all the reference individuals.

SNP based on highest MAF provided the second highest correlations after the (co)variance method. Random selection of SNP at lowest densities provided the poorest accuracies with little difference between RANI and RANC. As more SNP were selected, random selection of SNP started to outperform selection based on MAF such that at densities higher than 20K, RANI and RANC always provided higher accuracies than MAFI and MAFC, though the differences were marginal. The standard errors of imputation accuracies (result not shown) were always smaller than 0.02 with lower values at higher densities.

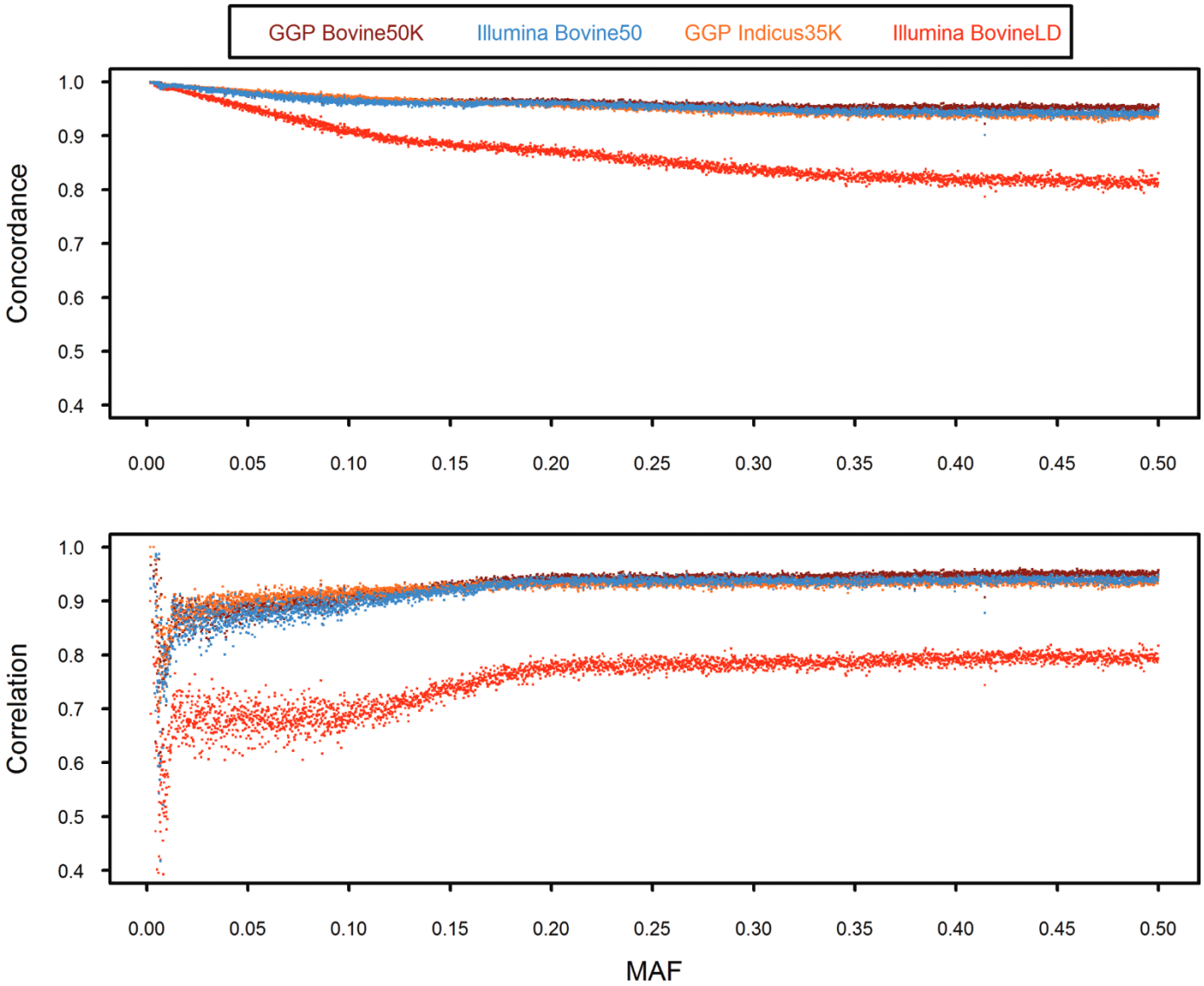

Figure 3. Concordance values (top) and correlations (bottom) of individual SNP against their minor allele frequency (MAF) obtained from the optimal imputation of different commercial arrays. GeneSeek-Genomic-Profiler (GGP) Bovine 50K and GGP Indicus 35K (Neogen Corporation, Lincoln, NE); Illumina Bovine50 and Illumina BovineLD (Illumina, San Diego, CA).

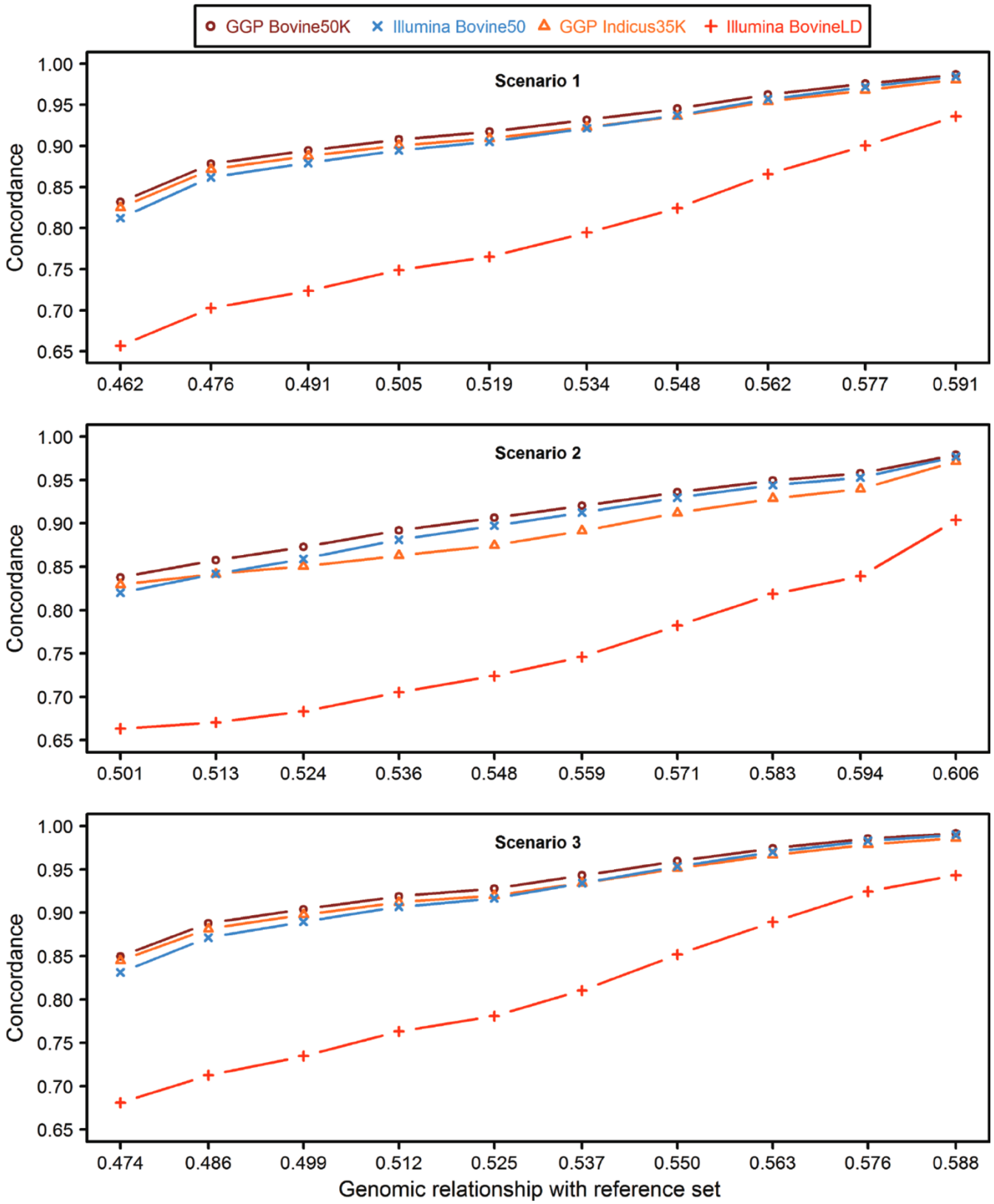

Figure 4. Concordance values of imputed genotypes of crossbred animals against their relationship with the reference set obtained from Minimac in different scenarios of imputation. GeneSeek-Genomic-Profiler (GGP) Bovine 50K and GGP Indicus 35K (Neogen Corporation, Lincoln, NE); Illumina Bovine50 and Illumina BovineLD (Illumina, San Diego, CA). Color version available online.

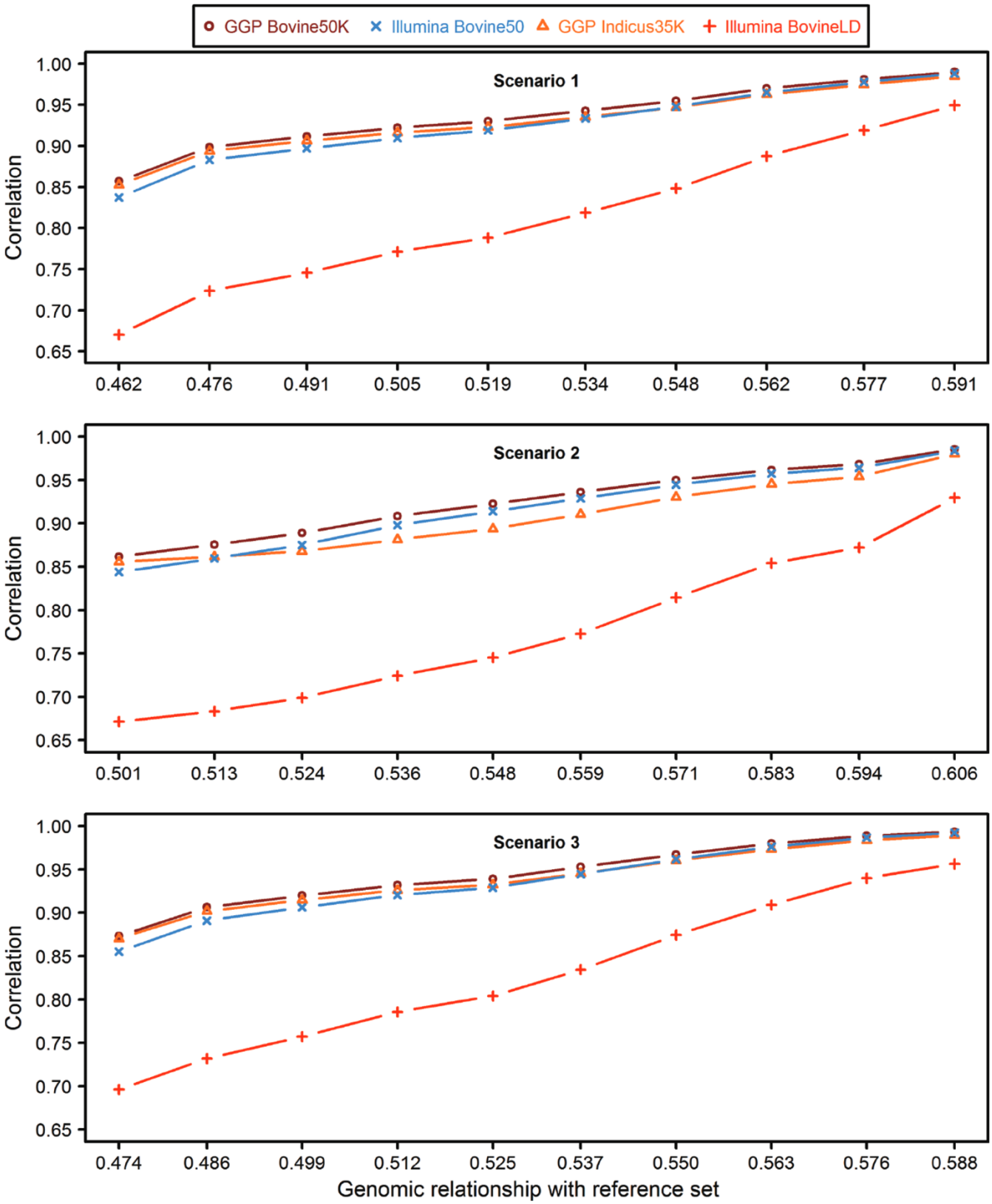

Figure 5. Correlations between real and imputed genotypes of crossbred animals against their relationship with the reference set obtained from Minimac in different scenarios of imputation. GeneSeek-Genomic-Profiler (GGP) Bovine 50K and GGP Indicus 35K (Neogen Corporation, Lincoln, NE); Illumina Bovine50 and Illumina BovineLD (Illumina, San Diego, CA). Color version available online.

Genomic Predictions Using Reduced and Imputed SNP Panels

The relative accuracies of genomic predictions obtained from the reduced sets of SNP and also from the imputed genotypes are in Table 7 for different SNP selection methods and different thresholds placed for the variance explained. Genomic predictions based on SNP selected by the (co)variance method achieved higher accuracies compared with other methods of selecting SNP, especially when less SNP were selected. When the threshold on variance explained was at 40% and higher, RANI and RANC achieved marginally higher genomic prediction accuracies compared with other methods of SNP selection, but all methods gave r^2 higher than 0.99. Accuracies of genomic prediction were always higher when the reduced SNP were imputed to the HD panel and then used to estimate GEBV rather than being used directly in genomic evaluations. The prediction accuracies from imputed genotypes were consistent with but substantially higher than accuracies achieved in imputations. Using the best method of SNP selection (i.e., COV), an imputation accuracy of 0.64 was achieved when 3.8K SNP were selected and this subsequently gave an accuracy of 0.95 for GEBV. Using the same selected SNP in genomic evaluation without imputation gave a substantially lower accuracy of 0.70 for GEBV.

DISCUSSION

The results of the current study confirm that genotype imputation can be successfully applied in crossbred dairy cattle populations of East Africa. An imputation scenario including both crossbred and ancestral purebred animals in the reference set and using Minimac as the imputation software achieved the highest accuracy in imputation of commercial SNP panels and it was regarded as the optimal imputation strategy. The accuracy of imputation and of GEBV from reduced SNP assays was higher when low density SNP were selected using a method based on the (co)variance between SNP and weighted by their MAF. At higher SNP densities, however, no significant differences were present between accuracies of imputations or GEBV from different SNP selection methods. The presented method can be used as a guide to select the required number of SNP from different regions of genome based on the varying structure of LD between markers, to design low-density arrays optimized for imputation to higher densities.

Table 6. The number of selected SNP at different thresholds for the proportion of variance explained and imputation accuracies obtained from different methods of SNP selection

Variance explained (%)	No. of SNP[1]	Imputation accuracy from SNP selection methods[1,2]									
		COV		MAFI		RANI		RANC		MAFC	
		Con	Cor	Con	Cor	Con	Cor	Con	Cor	Con	Cor
1	3,757	0.7905	0.6379	0.7733	0.6077	0.6970	0.4611	0.6974	0.4680	0.7024	0.4886
5	4,013	0.8109	0.6778	0.7930	0.6458	0.7134	0.4920	0.7194	0.5094	0.7244	0.5225
7	4,725	0.8393	0.7316	0.8218	0.7020	0.7695	0.6072	0.7646	0.5961	0.7716	0.6086
10	6,166	0.8684	0.7834	0.8511	0.7553	0.8263	0.7166	0.8204	0.7033	0.8217	0.7005
12	7,162	0.8810	0.8049	0.8636	0.7772	0.8462	0.7522	0.8422	0.7431	0.8413	0.7361
15	8,738	0.8954	0.8292	0.8778	0.8016	0.8663	0.7871	0.8632	0.7803	0.8606	0.7701
20	11,773	0.9140	0.8599	0.8963	0.8325	0.8907	0.8281	0.8893	0.8245	0.8838	0.8101
25	15,373	0.9283	0.8832	0.9107	0.8561	0.9087	0.8578	0.9078	0.8552	0.9008	0.8385
30	19,812	0.9398	0.9017	0.9228	0.8756	0.9235	0.8816	0.9232	0.8802	0.9155	0.8626
35	25,410	0.9495	0.9170	0.9336	0.8928	0.9365	0.9021	0.9364	0.9015	0.9281	0.8830
40	32,573	0.9576	0.9299	0.9437	0.9088	0.9480	0.9203	0.9482	0.9199	0.9394	0.9010
45	41,383	0.9644	0.9405	0.9525	0.9226	0.9579	0.9355	0.9579	0.9351	0.9490	0.9161
50	52,134	0.9700	0.9495	0.9602	0.9344	0.9661	0.9481	0.9661	0.9478	0.9573	0.9291
55	64,907	0.9747	0.9567	0.9666	0.9444	0.9729	0.9585	0.9727	0.9579	0.9643	0.9401
60	79,831	0.9787	0.9630	0.9720	0.9526	0.9782	0.9665	0.9778	0.9658	0.9701	0.9490
65	97,613	0.9823	0.9686	0.9765	0.9596	0.9825	0.9731	0.9820	0.9722	0.9748	0.9563
70	119,120	0.9853	0.9734	0.9805	0.9657	0.9859	0.9783	0.9855	0.9775	0.9791	0.9629
75	144,995	0.9880	0.9776	0.9840	0.9710	0.9887	0.9825	0.9883	0.9818	0.9827	0.9684
80	177,382	0.9904	0.9814	0.9869	0.9756	0.9909	0.9859	0.9905	0.9853	0.9858	0.9733
85	220,109	0.9926	0.9848	0.9896	0.9797	0.9928	0.9888	0.9924	0.9882	0.9887	0.9777
90	281,030	0.9947	0.9881	0.9922	0.9836	0.9944	0.9913	0.9941	0.9907	0.9915	0.9819
95	378,216	0.9969	0.9917	0.9946	0.9869	0.9958	0.9935	0.9955	0.9930	0.9942	0.9857

[1]Averaged across 5 folds.

[2]Proportion of correctly imputed genotypes (Con) and average correlations between real and imputed genotypes (Cor) obtained from imputation of subsets of selected SNP based on COV = (co)variance method; MAFI = minor allele frequency within interval; RANI = random within interval; RANC = random across chromosome; and MAFC = minor allele frequency across chromosome.

The accuracy of imputation with low density arrays largely depends on the strategy used to impute genotypes. This includes the choice of reference population and imputation algorithm among other factors influencing the imputation accuracy. It has been shown that a larger size of the reference population generally increases the imputation accuracy (e.g., Hozé et al., 2013). However, having immediate ancestors of the target individuals in the reference set is more important than the sample size (e.g., Ventura et al., 2014). Crossbred dairy cattle in East Africa form an admixed population of animals with different breed proportions from various indigenous and exotic dairy breeds (Strucken et al., 2017). To achieve a high level of imputation accuracy, therefore, it is necessary to create a reference set with high relationships to the crossbred animals and also an optimal contribution to the breed composition of crossbred animals. In the present study, a mixture of animals from crossbred, indigenous, and exotic groups when used as the reference set (i.e., scenario 3) resulted in higher top relationships between target and reference populations (Table 5) and consequently higher accuracies of imputation compared with other scenarios (Tables 3 and 4). The inclusion of crossbreds in the reference set provides closely related individuals to the target set, which allows high imputation accuracies (scenario 1). Adding to the imputation process purebred indigenous and exotic animals representative of the ancestral breeds contributing to the imputed crossbred animals improves the accuracy of imputation only marginally (scenario 3 vs. scenario 1). The crossbred dairy cattle in East Africa result from many generations of crossing of animals with different breed proportions, and the individual purebred animals used as reference in imputation are not themselves ancestors of the crossbred animals. Thus the long-range haplotypes in crossbreds will likely differ from those in reference purebreds, explaining why use of just the reference purebreds in imputation (i.e., scenario 2) performed poorly. Similar to our findings, Ventura et al. (2016) reported that a large reference population including all the available data from all breeds is preferred over smaller within breed reference sets in imputation of multi-breed sheep populations of New Zealand.

A positive relationship between relatedness to the reference set and imputation accuracy was observed in all scenarios of imputation (Figures 4 and 5). This confirms the importance of maintaining a high level of connectedness between reference and target animals to achieve a high imputation accuracy. The relationship between connectedness to the reference set and the realized imputation accuracy can also be used to

Table 7. Squared correlations between genomic breeding values estimated using the real high-density panel and those from reduced (Sel) and imputed (Imp) genotypes at different SNP densities selected by different methods of SNP selection

Variance explained (%)	No. of SNP[1]	Genomic prediction accuracy from SNP selection methods[1,2]									
		COV		MAFI		RANI		RANC		MAFC	
		Sel	Imp	Sel	Imp	Sel	Imp	Sel	Imp	Sel	Imp
1	3,757	0.6990	0.9474	0.6520	0.9279	0.6775	0.8669	0.6592	0.8706	0.5854	0.8710
5	4,013	0.7102	0.9558	0.6485	0.9398	0.6761	0.8890	0.6512	0.8979	0.5855	0.8932
7	4,725	0.7364	0.9666	0.6635	0.9536	0.6771	0.9266	0.6860	0.9242	0.6009	0.9196
10	6,166	0.7778	0.9770	0.6920	0.9662	0.7457	0.9579	0.7114	0.9571	0.6468	0.9489
12	7,162	0.8017	0.9803	0.7159	0.9698	0.7438	0.9678	0.7440	0.9661	0.6619	0.9608
15	8,738	0.8250	0.9849	0.7393	0.9763	0.7907	0.9732	0.7911	0.9740	0.7017	0.9688
20	11,773	0.8547	0.9907	0.7886	0.9830	0.8183	0.9811	0.8311	0.9825	0.7421	0.9793
25	15,373	0.8748	0.9929	0.8166	0.9870	0.8282	0.9864	0.8568	0.9873	0.7726	0.9847
30	19,812	0.8907	0.9949	0.8421	0.9901	0.8817	0.9907	0.8839	0.9902	0.8049	0.9885
35	25,410	0.9072	0.9961	0.8598	0.9927	0.8993	0.9931	0.9018	0.9931	0.8366	0.9915
40	32,573	0.9161	0.9967	0.8797	0.9944	0.9188	0.9949	0.9295	0.9956	0.8620	0.9945
45	41,383	0.9251	0.9976	0.8956	0.9955	0.9339	0.9964	0.9385	0.9970	0.8856	0.9957
50	52,134	0.9320	0.9982	0.9062	0.9967	0.9463	0.9977	0.9515	0.9980	0.9048	0.9966
55	64,907	0.9400	0.9986	0.9177	0.9975	0.9558	0.9984	0.9611	0.9985	0.9157	0.9977
60	79,831	0.9466	0.9989	0.9295	0.9982	0.9603	0.9990	0.9696	0.9990	0.9273	0.9982
65	97,613	0.9528	0.9992	0.9396	0.9987	0.9690	0.9993	0.9761	0.9993	0.9358	0.9986
70	119,120	0.9610	0.9994	0.9467	0.9989	0.9758	0.9995	0.9805	0.9995	0.9449	0.9990
75	144,995	0.9676	0.9996	0.9564	0.9993	0.9793	0.9997	0.9850	0.9997	0.9541	0.9992
80	177,382	0.9723	0.9998	0.9628	0.9996	0.9827	0.9998	0.9885	0.9998	0.9626	0.9995
85	220,109	0.9787	0.9999	0.9704	0.9997	0.9876	0.9999	0.9912	0.9999	0.9702	0.9997
90	281,030	0.9855	0.9999	0.9795	0.9999	0.9905	0.9999	0.9937	0.9999	0.9794	0.9998
95	378,216	0.9927	1	0.9895	0.9999	0.9942	1	0.9966	1	0.9901	0.9999

[1]Averaged across 5 folds.

[2]Selection of SNP based on COV = (co)variance method; MAFI = minor allele frequency within interval; RANI = random within interval; RANC = random across chromosome; and MAFC = minor allele frequency across chromosome.

predict an expected imputation accuracy for target animals before undertaking an imputation (e.g., Ventura et al., 2016). Genomic relationships between different populations can be similarly used to predict accuracy of across-population imputations. Here we observed that the accuracies of across-population imputations were related to the average genomic relationships between the populations (Supplemental Tables S1 to S2; https://doi.org/10.3168/jds.2018-14621).

In addition to the structure of reference population, the imputation algorithm can affect the imputation results. Several imputation algorithms are available that either use the LD in population [population imputation: e.g., Beagle (Browning and Browning, 2016); Minimac (Howie et al., 2012)] or additionally incorporate family relationships between individuals [family imputation: e.g., Alphaimpute (Hickey et al., 2012b); FImpute (Sargolzaei et al., 2014)] to implement imputation. Several studies have compared the performance of different imputation software in livestock species and found that different methods are preferred in different situations (e.g., Hayes et al., 2012; Khatkar et al., 2012). In pedigreed populations, the incorporation of family information allows the tracking of long haplotypes that run within families and have low frequency in the population. This is particularly beneficial in imputation of rare alleles, which have a low imputation accuracy from population-based imputations (Sargolzaei et al., 2014). The observed trend between the MAF of SNP and the 2 measures of accuracy of imputation used in this study (Figure 3) has been reported in previous studies (e.g., Hayes et al., 2012; Hickey et al., 2012a) and the reasons behind these trends has been documented (Calus et al., 2014). The SNP with low MAF show a high concordance value (Figure 3) simply because there is a higher chance of correctly inferring rare alleles based on population allele frequencies by assigning the major allele as the missing allele. Correlation accounts for the MAF of SNP so it can be used to compare imputation accuracy across different loci with different MAF (Calus et al., 2014).

The accuracy of imputation for SNP which have specific applications in genetic studies (e.g., parentage tests and calculation of breed composition) is of additional interest. Using a similar data, Strucken et al. (2017) showed that SNP with largest allele frequency difference between European dairy breeds and a combined Nelore plus N'Dama population (as representatives of the ancient ancestors of indigenous East African zebu cattle) are the most informative for calculation of breed proportions of East African crossbred animals. Here, we compared the imputation accuracy of SNP grouped according to the difference in average allele frequencies between indigenous versus exotic dairy breeds (Figure 6). We found that accuracies were highest for loci with high allele frequency difference between the ancestral breeds, confirming that the imputed genotypes can be safely used for calculation of breed proportions of crossbred animals. These loci are also the most powerful SNP for undertaking genome-wide association studies to detect regions of genome that cause the very large differences in performance and adaptation traits of exotic dairy versus indigenous breeds. Because such loci are very poorly represented on all commercial assays other than the Illumina BovineHD BeadChip, this property of the imputation is particularly valuable.

The sensitivity of the imputation algorithm to size of the reference population and the density of the imputed and reference arrays should also be considered. In our study, Beagle gave the lowest imputation accuracies in scenario 1 compared with FImpute and Minimac. The inclusion of purebred in addition to crossbreds in the reference (i.e., scenario 3) increased the accuracies from Beagle especially at higher SNP densities but only slightly improved the accuracies for FImpute and Minimac (Tables 3 and 4). Although this could be partly due to the change in the structure of reference population (as discussed above), it can also be related to the increase in the reference sample size, which could imply that Beagle is more dependent on the size of reference population, as reported by Ventura et al. (2014). On the other hand, no increase occurred in imputation accuracy using Beagle from scenario 1 to scenario 3 for the Illumina BovineLD which could also indicate that the accuracy from Beagle is sensitive to the density of imputed panel. Similar results were reported by Sargolzaei et al. (2014) who showed that the imputation from lower densities using Beagle is more dependent on the size of reference set.

Cost-effective genotyping requires an appropriate balance between the size of reference and target populations. In an additional analysis, we examined the effect of having different proportions of reference and target animals in imputation. We found that when only 10% of the population (≈308 animals) was used as the reference to impute the remaining 90% as the target, the accuracy of imputation was still reasonably high. The increase in imputation accuracy was not proportional to the increase in reference sample size (Supplemental Table S3; https://doi.org/10.3168/jds.2018-14621), with only modest increase in accuracy especially when more than 30% of animals were in the reference set. As long as the reference set is a representative sample of the target population, the imputation accuracy should be related to the number of animals rather than the proportion of animals in the reference set. That is, the current reference set could be used to impute with the same accuracy as observed here the genotypes of an in-

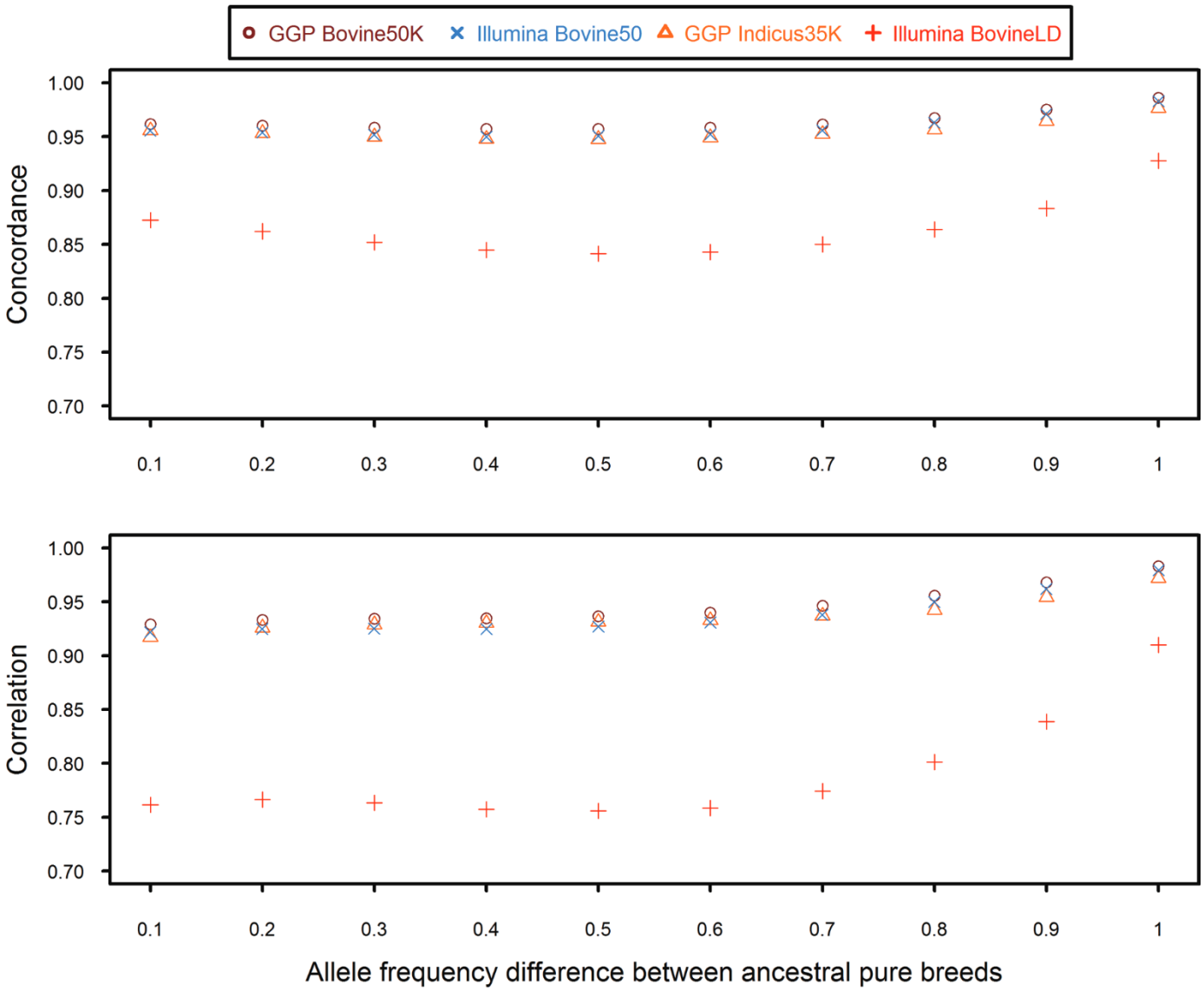

Figure 6. Average concordance values (top) and correlations (bottom) of SNP obtained from optimal imputation of crossbred animals and grouped based on allele frequency difference between ancestral pure breeds. GeneSeek-Genomic-Profiler (GGP) Bovine 50K and GGP Indicus 35K (Neogen Corporation, Lincoln, NE); Illumina Bovine50 and Illumina BovineLD (Illumina, San Diego, CA). Color version available online.

finite number of animals sampled from the same population in future. The relationship of sampled animals in future to the current reference set will provide a test of whether the reference set is properly representative of newly sampled animals.

In addition to the factors discussed above, the accuracy of imputation depends on the characteristics of SNP included in the low-density panel. It is expected that if SNP in the low-density panel have high LD with the missing genotypes, they should achieve a high imputation accuracy, especially in population imputation methods. But, if markers in the low-density panel are in strong LD with each other, they will provide redundant information for imputation. It is also known that SNP with high MAF have more information content for imputation (Hayes et al., 2012). The (co)variance method for selecting SNP presented in the current study is formulated to maximize LD with missing loci while minimizing LD among SNP on the assay and maintaining high MAF.

We applied our SNP selection method to select SNP within overlapping intervals of an arbitrary length of 1 Mbp across the whole genome, which was equivalent to a block with approximately 1% recombination per generation. Ideally, SNP should be selected within LD blocks separated by recombination hotspots. This

would minimize the crossing over that weakens the LD between markers so that the selected markers remain useful for imputation over many generations. However, there is currently no high accuracy recombination map for the bovine HD genotypes that can be used to define SNP selection intervals. Alternatively, given that LD varies across the bovine genome (e.g., Sargolzaei et al., 2008), the interval for selecting markers could be also varied to match the observed pattern of LD. The (co) variance method will account for the variation in LD across the genome because it selects SNP based on accounting for a target proportion of variation in SNP. More SNP will be selected in regions where LD is low than where LD is high.

The selection of SNP could be undertaken within larger intervals or even across the whole chromosome. This would decrease the number of selected SNP that is required to explain a certain amount of variation between SNP compared with choosing SNP from shorter intervals. But selection of SNP from long intervals would lead to selections based on lower levels of LD that are less useful for imputation and more likely to decay over time than selection from short intervals (de Roos et al., 2008). Genomic intervals could be defined so that the average LD between markers within the interval pass a predefined threshold. For example, Meuwissen et al. (2001) suggested that a squared correlation between adjacent markers equal to 0.2 or 0.3 is useful for GS. The selection of SNP within small intervals will be particularly important in crossbred populations such as we examined where there is no pedigree information, very little family structure, many generations of recombination that break up ancestral haplotypes, and low levels of LD (Figure 1 and Supplemental Table S4; https://doi.org/10.3168/jds.2018-14621) and hence large effective population size.

The (co)variance method for SNP selection provided the highest accuracy of imputation of low-density genotypes of East African crossbred dairy cattle compared with other methods tested in this study (Table 6). Selection of SNP based on MAFI was inferior to the COV method because it does not account for the LD between SNP and hence can select subsets of SNP that have high LD with each other and generally contain lower information. This problem is worse when SNP are selected based on highest MAF across the whole chromosome (MAFC) because MAFC is not optimized for uniformity across the chromosome and hence can leave gaps with little information for imputation. Random selection of SNP within intervals (RANI) or across chromosomes (RANC) provided very similar accuracies to each other at all densities. This suggests that even at the lowest densities used here, uniformity of marker spacing is not a particularly important factor for accuracy of imputation if SNP are selected at random.

The imputation accuracies in Table 6 are somewhat lower than accuracies reported in the literature for purebred dairy populations but are within the same range of those from populations with greater genetic diversity (e.g., Hozé et al., 2013). Crossbred populations resulting from many generations of admixture are expected to have larger N_e and weaker long-distance LD compared with purebred populations (e.g., Lu et al., 2012). Hence, larger reference populations are required to capture the haplotype diversity in crossbreeds and to achieve a similar accuracy of imputation to those in purebreds. Hozé et al. (2013) reported lower imputation accuracies in beef breeds compared with dairy breeds where the former group in general showed a higher rate of decay of LD across their genome. Bolormaa et al. (2015) also reported lower imputation accuracies for a crossbred sheep population than those obtained for purebred sheep breeds.

The squared correlations (r^2) between $GEBV_{HD}$ and $GEBV_{IMP}$ shown in Table 7 confirm that the genotype imputation is an effective approach for obtaining accurate GEBV from low-density genotypes in the crossbred dairy cattle populations of East Africa. The r^2 between $GEBV_{HD}$ and $GEBV_{IMP}$ was high even when small number of SNP were used in imputation and the accuracy of imputation was lowest. The high r^2 obtained from $GEBV_{IMP}$ also implies that the loss in genetic gain from using imputed genotypes instead of real genotypes is minimal in East African crossbred dairy cattle. Imputation will increase the accuracy of GEBV further through allowing a bigger reference population for a given genotyping budget. We obtained an r^2 of 0.95 between $GEBV_{HD}$ and $GEBV_{IMP}$ when around 4K SNP were imputed up to 700K HD genotypes. Similar patterns to results of the current study have been reported in previous studies. Weigel et al. (2010) showed that the predictive ability of imputed genotypes is 95% of that from the real genotypes when 3K genotypes are imputed up to 50K in Jersey cattle. Daetwyler et al. (2011) found that the accuracy of GEBV from imputed genotypes was 95% of that from real genotypes when 87.8% of genotypes were correctly imputed. Cleveland and Hickey (2013) reported a correlation of 0.95 between GEBV from genotypes imputed from 3K to 50K compared with the real 50K genotypes in pigs.

The r^2 between $GEBV_{HD}$ and $GEBV_{SEL}$, however, was small when low-density SNP panels were directly used in estimation of GEBV without imputation and it only went above 0.95 when at least 80K SNP were included. Therefore, the best strategy in obtaining GEBV from low-density panels is to first impute them to HD geno-

types and then incorporate the imputed genotypes in GS. Even when the imputations have high error rates, the bias from the wrongly inferred genotypes will not propagate in accuracy of subsequent genomic prediction (Wu et al., 2016) and the effectiveness of selection on $GEBV_{IMP}$ would be very similar to selection based on $GEBV_{HD}$.

The (co)variance method presented in the current study was optimized to achieve a high accuracy of genotype imputation, yet it can also be used to select reduced SNP assays for GS. Since GS relies on LD between markers and QTL to estimate GEBV, the selection of SNP for direct use in GS might not need an adjustment for MAF. To compare genomic prediction accuracies obtained from SNP selected with ($w = 1$) and without ($w = 0$) adjustment for MAF, we also selected different densities of SNP with no adjustment for MAF and subsequently used them for prediction of GEBV. The accuracy of genomic prediction from (co)variance method with no adjustment for MAF was highest at almost all densities of selected SNP compared with other SNP selection method (Supplemental Table S5; https://doi.org/10.3168/jds.2018-14621). The accuracies from the (co)variance method with an adjustment for MAF ($w = 1$) were higher when less than around 15K SNP were selected but higher accuracies obtained with no adjustment for MAF ($w = 0$) at higher densities of selected SNP. Genomic prediction methods that rely on information from individual SNP genotypes rather than a genomic relationship matrix (e.g., Bayesian methods) might achieve higher accuracies of genomic prediction when SNP from the (co)variance method with no adjustment for MAF are used to predict GEBV.

CONCLUSIONS

The results of the current study confirm that the genotypes of East African crossbred dairy cattle can be imputed with sufficiently high accuracy to achieve highly accurate GEBV, thus providing large savings in genotyping costs, or increased effectiveness of selection for a fixed genotyping budget. These results will be applied to optimize genotyping within the current Africa Dairy Genetic Gains program, which aims to demonstrate effective and sustainable genetic improvement for smallholder crossbred cattle in East Africa. There is no great advantage of including the current purebred reference animals in the crossbred imputation, but if the actual purebred ancestors were available, using them for imputation of crossbred animals might provide a greater increase in accuracy. The presented method for SNP selection is straightforward in its application and can be useful in any population, especially those that rely on population-based methods of imputation.

ACKNOWLEDGMENTS

The authors thank Bill & Melinda Gates Foundation for funding the Dairy Genetics East Africa (DGEA) project. Illumina (Illumina, San Diego, CA) and Geneseek (Neogen Corporation, Lincoln, NE) kindly provided contributions to genotyping costs. Special thanks to Ed Rege (PICO Eastern Africa, Nairobi, Kenya) who co-designed and helped leading the DGEA project, and to Julie Ojango, James Rao, Denis Mujibi, and Tadelle Dessie of International Livestock Research Institute (ILRI, Kenya and Ethiopia) who facilitated and undertook the field sampling that allowed this research to be undertaken. The British Friesian genotype data was kindly provided by Scottish Rural University College (SRUC, Scotland), and the Ayrshire genotypes were kindly supplied by the Canadian Dairy Network (CDN, Canada). We also thank the smallholder farmers who participated in the DGEA project and provided samples and data on their animals.

REFERENCES

Boichard, D., H. Chung, R. Dassonneville, X. David, A. Eggen, S. Fritz, K. J. Gietzen, B. J. Hayes, C. T. Lawley, T. S. Sonstegard, C. P. Van Tassell, P. M. VanRaden, K. A. Viaud-Martinez, G. R. Wiggans, and Bovine LD Consortium. 2012. Design of a bovine low-density SNP array optimized for imputation. PLoS One 7:e34130.

Bolormaa, S., K. Gore, J. H. J. van der Werf, B. J. Hayes, and H. D. Daetwyler. 2015. Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. Anim. Genet. 46:544–556.

Browning, B. L., and S. R. Browning. 2016. Genotype imputation with millions of reference samples. Am. J. Hum. Genet. 98:116–126.

Calus, M. P. L., A. C. Bouwman, J. M. Hickey, R. F. Veerkamp, and H. A. Mulder. 2014. Evaluation of measures of correctness of genotype imputation in the context of genomic prediction: A review of livestock applications. Animal 8:1743–1753.

Cleveland, M. A., and J. M. Hickey. 2013. Practical implementation of cost-effective genomic selection in commercial pig breeding using imputation. J. Anim. Sci. 91:3583–3592.

Corbin, L. J., A. Kranis, S. C. Blott, J. E. Swinburne, M. Vaudin, S. C. Bishop, and J. A. Woolliams. 2014. The utility of low-density genotyping for imputation in the Thoroughbred horse. Genet. Sel. Evol. 46:9.

Daetwyler, H. D., G. R. Wiggans, B. J. Hayes, J. A. Woolliams, and M. E. Goddard. 2011. Imputation of missing genotypes from sparse to high density using long-range phasing. Genetics 189:317–327.

Das, S., L. Forer, S. Schönherr, C. Sidore, A. E. Locke, A. Kwong, S. I. Vrieze, E. Y. Chew, S. Levy, M. McGue, D. Schlessinger, D. Stambolian, P.-R. Loh, W. G. Iacono, A. Swaroop, L. J. Scott, F. Cucca, F. Kronenberg, M. Boehnke, G. R. Abecasis, and C. Fuchsberger. 2016. Next-generation genotype imputation service and methods. Nat. Genet. 48:1284.

de Roos, A. P. W., B. J. Hayes, R. J. Spelman, and M. E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holstein–Friesian, Jersey and Angus cattle. Genetics 179:1503–1512.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2009. Genomic selection using low-density marker panels. Genetics 182:343–353.

Hayes, B. J., P. J. Bowman, H. D. Daetwyler, J. W. Kijas, and J. H. J. van der Werf. 2012. Accuracy of genotype imputation in sheep breeds. Anim. Genet. 43:72–80.

Hickey, J. M., J. Crossa, R. Babu, and G. de los Campos. 2012a. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. Crop Sci. 52:654–663.

Hickey, J. M., B. P. Kinghorn, B. Tier, J. H. van der Werf, and M. A. Cleveland. 2012b. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet. Sel. Evol. 44:9.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44:955.

Hozé, C., M.-N. Fouilloux, E. Venot, F. Guillaume, R. Dassonneville, S. Fritz, V. Ducrocq, F. Phocas, D. Boichard, and P. Croiseau. 2013. High-density marker imputation accuracy in sixteen French cattle breeds. Genet. Sel. Evol. 45:33.

Khatkar, M. S., G. Moser, B. J. Hayes, and H. W. Raadsma. 2012. Strategies and utility of imputed SNP genotypes for genomic analysis in dairy cattle. BMC Genomics 13:538.

Loh, P.-R., P. F. Palamara, and A. L. Price. 2016. Fast and accurate long-range phasing in a UK Biobank cohort. Nat. Genet. 48:811.

Lu, D., M. Sargolzaei, M. Kelly, C. Li, G. Vander Voort, Z. Wang, G. Plastow, S. Moore, and S. Miller. 2012. Linkage disequilibrium in Angus, Charolais, and Crossbred beef cattle. Front. Genet. 3.

Mathew, B., J. Léon, and M. J. Sillanpää. 2018. A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction. Heredity 120:356–368.

Meuwissen, T., B. Hayes, and M. Goddard. 2016. Genomic selection: A paradigm shift in animal breeding. Anim. Front. 6:6–14.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Ojango, J. M. K., A. Marete, D. Mujibi, J. Rao, J. Pool, J. E. O. Rege, C. Gondro, W. M. S. P. Weerasinghe, J. P. Gibson, and A. M. Okeyo. 2014. A novel use of high density SNP assays to optimize choice of different crossbred dairy cattle genotypes in small-holder systems in East Africa. Pages 2–4 in Proc. 10th World Congr. Genet. Appl. to Livest. Prod. Am. Soc. Anim. Sci., Champaign, IL.

Sargolzaei, M., J. P. Chesnais, and F. S. Schenkel. 2014. A new approach for efficient genotype imputation using information from relatives. BMC Genomics 15:478.

Sargolzaei, M., F. S. Schenkel, G. B. Jansen, and L. R. Schaeffer. 2008. Extent of linkage disequilibrium in Holstein cattle in North America. J. Dairy Sci. 91:2106–2117.

Strucken, E. M., H. A. Al-Mamun, C. Esquivelzeta-Rabell, C. Gondro, O. A. Mwai, and J. P. Gibson. 2017. Genetic tests for estimating dairy breed proportion and parentage assignment in East African crossbred cattle. Genet. Sel. Evol. 49:67.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91:4414–4423.

Ventura, R. V., D. Lu, F. S. Schenkel, Z. Wang, C. Li, and S. P. Miller. 2014. Impact of reference population on accuracy of imputation from 6K to 50K single nucleotide polymorphism chips in purebred and crossbreed beef cattle1. J. Anim. Sci. 92:1433–1444.

Ventura, R. V., S. P. Miller, K. G. Dodds, B. Auvray, M. Lee, M. Bixley, S. M. Clarke, and J. C. McEwan. 2016. Assessing accuracy of imputation using different SNP panel densities in a multi-breed sheep population. Genet. Sel. Evol. 48:71.

Weigel, K. A., G. de los Campos, A. I. Vazquez, G. J. M. Rosa, D. Gianola, and C. P. Van Tassell. 2010. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. J. Dairy Sci. 93:5423–5435.

Wu, X.-L., J. Xu, G. Feng, G. R. Wiggans, J. F. Taylor, J. He, C. Qian, J. Qiu, B. Simpson, J. Walker, and S. Bauck. 2016. Optimal design of low-density SNP arrays for genomic prediction: Algorithm and applications. PLoS One 11:e0161719.

Zimin, A. V., A. L. Delcher, L. Florea, D. R. Kelley, M. C. Schatz, D. Puiu, F. Hanrahan, G. Pertea, C. P. Van Tassell, T. S. Sonstegard, G. Marçais, M. Roberts, P. Subramanian, J. A. Yorke, and S. L. Salzberg. 2009. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol. 10:R42.